Invariant Feature Subspace Recovery for Multi-Class Classification

Gargi Balasubramaniam
University of Illinois, Urbana Champaign
gargib2@illinois.edu

Haoxiang Wang
University of Illinois, Urbana Champaign
hwang264@illinois.edu

Han Zhao
University of Illinois, Urbana Champaign
hanzhao@illinois.edu

Abstract

Domain generalization aims to learn a model over multiple training environments to generalize to unseen environments. Recently, Wang et al. [2022] proposed Invariant-feature Subspace Recovery (ISR), a domain generalization algorithm that uses the means of class-conditional data distributions to provably identify the invariant-feature subspace under a given causal model. However, due to the specific assumptions of the causal model, the original ISR algorithm is conditioned on a single class only, without utilizing information from the rest of the classes. In this work, we consider the setting of multi-class classification under a more general causal model, and propose an extension of the ISR algorithm, called ISR-Multiclass. This proposed algorithm can provably recover the invariant-feature subspace with $\lceil d_{spu}/k \rceil + 1$ environments, where d_{spu} is the number of spurious features and k is the number of classes. Empirically, we first examine ISR-Multiclass in a synthetic dataset, and demonstrate its superiority over the original ISR in the multi-class setting. Furthermore, we conduct experiments in Multiclass Coloured MNIST, a semi-synthetic dataset with strong spurious correlations, and show that ISR-Multiclass can significantly improve the robustness of neural nets trained by various methods (e.g., ERM and IRM) against spurious correlations.

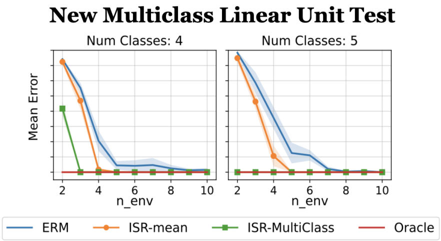
1 Introduction

Domain generalization involves a learner having access to several domains during training time, which can be leveraged together to generalize better to unseen domains during test time. Notably, a recent line of work (Wang et al. [2022], Rosenfeld et al. [2022], Kirichenko et al. [2022]) focuses on achieving robustness to spurious correlations via simple linear transformations/last layer re-training. In this work, we focus on and aim to extend the post-processing procedure proposed by Wang et al. [2022]. Wang et al. [2022] present the ISR-Mean[1], an algorithm that provably recovers the invariant feature subspace by using the 1^{st} order moments of the class conditional data distribution, under a causal model for classification considered by Rosenfeld et al. [2022], Arjovsky et al. [2019].

However, in the more realistic multi-class setting, it is unclear whether there exists any relation between the number of classes and the number of training environments required to recover invariant features. Further, Singla and Feizi [2021] demonstrates that features that are spurious for a given target label may be the core features for another. One could apply class-specific transformations to mitigate this. However, this is infeasible as we do not know the class labels during testing. In light of the above challenges, we are specifically motivated by the following question:

[1] Referred to as ISR in this paper.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).



Multiclass Coloured MNIST

Algorithm	Average Accuracy		Worst-10 Group Accuracy	
	Original	ISR-Multiclass	Original	ISR-Multiclass
ERM	58.20±1.03	**78.50±0.76**	2.35±0.59	**39.60±7.90**
IB-ERM	70.58±1.24	**81.40±1.15**	10.06±2.66	**42.63±6.77**
IRM	73.85±0.79	**82.01±0.97**	25.66±3.14	**45.36±6.46**
IB-IRM	77.81±0.84	**82.95±2.42**	32.14±2.48	**49.17±5.63**

Figure 1: ISR-Multiclass enables faster recovery (left) of invariant features and improves accuracy by large margin on coloured MNIST (right).

In the context of linear causal models, can the number of classes compensate for the number of training environments in recovering the invariant features?

In order to answer the above questions, we propose **ISR-Multiclass**, an extension of the ISR-Mean algorithm that provably recovers the invariant subspace in $\lceil d_{spu}/k \rceil + 1$ environments, where d_{spu} is the number of spurious features (i.e., the dimensionality of the spurious-feature subspace) and k is the number of classes. Note that our result improves over the original environmental complexity of $d_{spu} + 1$ for binary classification problem.[2], hence it provides an affirmative answer to the above problem. Furthermore, our result shows that the benefits of k-class classification problems help to reduce the environmental complexity by an order of k.

Additionally, we introduce a new multi-class benchmark as a linear unit test based on Aubin et al. [2021] and show that ISR-Multiclass can leverage class information to its advantage for faster recovery of invariant features. We also demonstrate significant improvements in performance on the multi-class Coloured MNIST dataset (Ahuja et al. [2021]. A summary is shown in Figure 1.

2 ISR Multiclass

In this section, we introduce ISR-Multiclass i.e. **I**nvariant Feature **S**ubspace **R**ecovery for **Multi Class** classification.

2.1 Setup

We study the causal model as shown in Figure 2, similar to that of Rosenfeld et al. [2020].

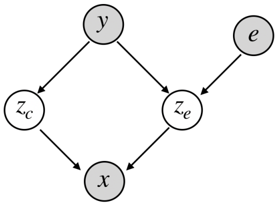


Figure 2: Data Model. Shading indicates that the variables are observed.

Let y be sampled from a prior distribution of labels $\{y_1, y_2, ..., y_k\}$. Let the dimension of the invariant and spurious features be d_c and d_s respectively, such that $d_c + d_s = d$, which is the total input dimension. Then, we sample:

$$z_c \sim \mathcal{N}(\mu_y, \sigma_c^2 I_{d_c}), z_e \sim \mathcal{N}(\mu_{ye}, \sigma_e^2 I_{d_s}) \tag{1}$$

[2] When $k = 2$, the environmental complexity in Wang et al. [2022] is $d_{spu} + 1$ instead of $d_{spu}/2 + 1$ because of one specific assumption on the symmetry of the conditional feature distributions, which we also remove in this work.

where $\mu_y \in \mathbb{R}^{d_c}, \mu_{ye} \in \mathbb{R}^{d_s}$ and $\sigma_c, \sigma_e \in \mathbb{R}_+$. As a comparison, in the original causal model for binary classification considered in Wang et al. [2022], Rosenfeld et al. [2020], the invariant feature distribution is given by $z_c \sim \mathcal{N}(y\mu_c, \sigma_c^2 I_{d_c})$, where $y \in \{+1, -1\}$. We point out that this is an unnecessary assumption on the causal model mainly imposed for technical convenience. Furthermore, as it will become clear shortly, this symmetry assumption on the means of the invariant feature distributions also artificially increases the environmental complexity from $d_{spu}/2 + 1$ to $d_{spu} + 1$.

Finally, the input x is generated from a linear transformation of the concatenated invariant and spurious features as follows:

$$x \leftarrow [A\ B] \begin{bmatrix} z_c \\ z_e \end{bmatrix} \tag{2}$$

$$= Az_c + Bz_e = Rz \in \mathbb{R}^d \tag{3}$$

As a consequence, the marginal distribution of x is the following:

$$\mathcal{N}(A\mu_y + B\mu_{ye}, \sigma_c^2 AA^T + \sigma_e^2 BB^T) \tag{4}$$

Here, $A = \mathbb{R}^{d \times d_c}, B = \mathbb{R}^{d \times d_s}$ are transformation matrices and $R = [A, B] \in \mathbb{R}^{d \times d}$.

In this work, we mainly focus on the design of the recovery algorithm itself, so for clarity of the derivation and discussion, we conveniently assume that the algorithm has access to infinite amount of data sampled according to the data model described above to limit the impact of finite samples, as a common practice also used in the prior works [Wang et al., 2022, Rosenfeld et al., 2020].

Similar to Wang et al. [2022], we adopt the full-rank assumption on the means of the spurious features from different environments and different classes.

Assumption 1 *For the set of environmental means,* $\{\mu_{ye}\}_{e=1,}^{E}{}_{y=1}^{k}$*, we assume that*

$$\dim(\mathrm{span}(\{\mu_{ye} : y \in [k], e \in [E]\})) = \min(E \times k, d_s) \tag{5}$$

Intuitively, this assumption ensures that the spurious (and thus invariant) subspace can be recovered and multiple classes / multiple environments do not *trivially replicate* information.

Further, we also assume the following as per Rosenfeld et al. [2020].

Assumption 2 R *is injective.*

2.2 ISR-Multiclass

We now present ISR-Multiclass, as outlined in Algorithm 1.The detailed steps can be found in Appendix A.1.

Briefly, consider the means $\bar{x}_{ek} = A\mu_k + B\mu_{ke}$ for a given class k and environment e. Then, let:

$$\mathcal{M}_k := \left[\bar{x}_{1k}^\top \cdots \bar{x}_{Ek}^\top\right]^\top \tag{6}$$

Eigendecomposition of each M_k results in $E - 1$ eigenvectors (from Assumption 1 and Wang et al. [2022]). Now, consider M_{total} defined as:

$$\mathcal{M}_{total} := [P_1 | P_2 | \cdots | P_k] \in \mathbb{R}^{d \times (E-1)k} \tag{7}$$

where P_i is the set of $E - 1$ eigenvectors obtained from the eigendecomposition of M_k. We now present Theorem 1 which formally states the proposed improvement. The proof can be found in Appendix A.2.

Theorem 1 (ISR-Multiclass) *Assume that* $E \geq \lceil d_s/k \rceil + 1$ *and we have infinite data samples from every environment. We perform **SVD** for* $\mathcal{M}_{total}$ *(defined in* (7)*), let* $\{\lambda_1, \ldots, \lambda_d\}$ *denote the set of singular values obtained in descending order. It is guaranteed that the top* d_s *singular values are non-zero, i.e.,*

$$\forall 1 \leq i \leq d_s,\ \lambda_i > 0$$

The eigenvectors corresponding to these top d_s *singular values, i.e.,* $\{P'_1, \ldots, P'_{d_s}\}$*, correspond to the* d_s *spurious dimensions. Consequently, the null space* $NullSpace([P'_1, \ldots, P'_{d_s}]^\top \in \mathbb{R}^{d_s \times d}) = [P''_1, \ldots, P''_{d_c}]^\top \in \mathbb{R}^{d_c \times d}$ *corresponds to the invariant dimension subspace. A classifier* f *fitted to training data transformed by* $x \mapsto [P''_1, \ldots, P''_{d_c}]^\top x$ *recovers the invariant optimal predictor.*

Algorithm 1 ISR-Multiclass

```
**Input:** Data of all training environments, $\{\mathcal{D}_e\}_{e\in[E]}$ across all classes $y \in \{y_1, y_2, \ldots, y_k\}$.
**for** $y = y_1, y_2, \ldots, y_k$ **do**
  **for** $e = 1, 2, \ldots, E$ **do**
    Estimate the sample mean of $\{x|(x, y) \in \mathcal{D}_e, y = y_k\}$ as $\bar{x}_{ke} \in \mathbb{R}^d$
  **end for**
**end for**
**1.** Construct $k$ matrices $\mathcal{M}_k \in \mathbb{R}^{E\times d}$ with the $e$-th row of the $k$-th matrix as $\bar{x}_{ke}^\top$ for $e \in [E]$
**2.** Apply PCA to each $\mathcal{M}_k$ to obtain set of eigenvectors. From these, chose eigenvectors corresponding to $E - 1$ highest eigenvalues to construct $\mathcal{P}_k := \left[P_{1k}|P_{2k}| \cdots |P_{(E-1)k}\right] \in \mathbb{R}^{d\times(E-1)}$
**3.** Stack each $P_k$ to obtain $\mathcal{M}_{total} := [P_1|P_2| \cdots |P_k] \in \mathbb{R}^{d\times(E-1)k}$
**4.** Apply SVD of $M_{total}$ to obtain eigenvectors $\{P'_1, ..., P'_d\}$ with eigenvalues $\{\lambda_1, ..., \lambda_d\}$
**5.** Stack $d_s$ eigenvectors with the highest eigenvalues to obtain a transformation matrix $P' \in \mathbb{R}^{d_s\times d}$
**6.** Take the null space of $P' \in \mathbb{R}^{d_s\times d}$ to obtain $P'' \in \mathbb{R}^{d_c\times d}$
**7.** Apply transformation $x \mapsto P''x$ on the training data and fit a linear classifier (with $w \in \mathbb{R}^{d_c}$, $b \in \mathbb{R}$)
Resulting predictor is $f(x) = w^\top P''x + b$
```

Optimality Our method involves applying an additional SVD operation over ISR-Mean, and thus the global optimality holds: a classifier trained on these features is globally optimal, similar to arguments in Wang et al. [2022].

Environment Complexity The environment complexity of ISR-Multiclass is $\lceil d_s/k \rceil + 1$ (detailed proof in A.2). The key observation here is that the column space of each P_i matrix for $i \in [k]$ only consists of the span of $\{\mu_{ie}\}_{e=1}^{E}$. So in order to identify the subspace of spurious features, one needs at least d_s linearly independent components. Hence, we only need to ensure that $(E - 1)k \geq d_s$ so that the column space of $\mathcal{M}_{total}$ is full rank. Solving this inequality leads us to the desired bound on the environmental complexity. This implies that we can leverage information from *multiple classes* to *reduce* the environmental complexity by a factor of $1/k$, as compared to $d_s + 1$ proposed in Wang et al. [2022], while only relying on the 1^{st} order moments of the data generating distribution. Intuitively, our method leverages *additional* information from multiple classes to find the common spurious feature subspace in lesser number of environments.

3 Experiments

We now present empirical improvements demonstrated by ISR-Multiclass on 2 datasets.

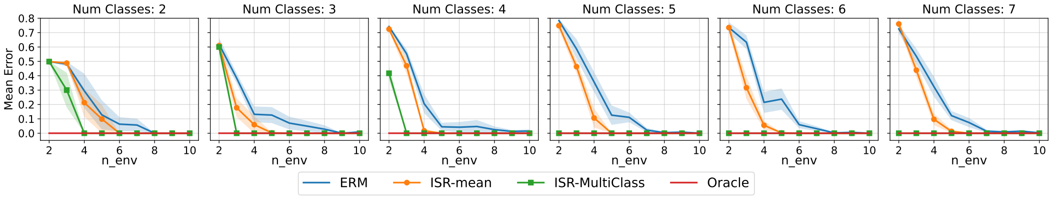


Figure 3: Evaluation on Multiclass Linear Unit Test Example 3s, where the y-axis denotes mean error over the test set. We have fixed $d_s = 5$ and $d_c = 5$, while k varies from 2 to 7. As indicated by our theoretical claim: ISR-Multiclass recovers features roughly in $\lceil d_s/k \rceil + 1$ environments. For example, when $k = 3$, we achieve optimal error in $\lceil 5/3 \rceil + 1 \approx 3$ environments. Similarly, beyond $k = 5$, we achieve optimality in 2 environments itself.

3.1 Multi-Class Linear Unit Test

We construct a multi-class version of Example 3 (and its scrabled version, Example 3s) as used in Aubin et al. [2021]. This is a synthetic dataset based on our causal model in section 2.1. Specific details of our construction can be found in Appendix B. Figure 3 depicts the improvement of ISR-

Multiclass compared to the original ISR-Mean [3] and ERM Vapnik [1991]. Evaluation is performed on the test split where the spurious dimensions are randomized. The Oracle is trained on this test split.

From figure 3, we observe that ISR-Multiclass is indeed able to leverage class information and recover invariant features to achieve optimal error with the number of environments inversely proportional to k, confirmed by our theoretical claim. Especially in the last three plots - with greater classes and lesser environments ($k > 5$ and $n_{env} < 5$), both ISR-Mean and ERM incur *higher* error, but ISR-Multiclass takes advantage of multiple classes to *improve* its performance instead and match the oracle.

3.2 Multi-Class Coloured MNIST (MC-CMNIST)

We consider the 10-class classification task of Coloured MNIST as proposed in Ahuja et al. [2021]. Note that in the train environments, every digit is highly correlated with a specific color, as depicted in Figure 1. This correlation breaks down in the test environment, i.e., every digit is randomly colored.

Evaluation Table 1 presents the results on MC-CMNIST. Performance is evaluated on every group, which denotes a specific combination of $(y, color) \in \mathcal{G} = \mathcal{Y} \times \mathcal{E}$. Note that there are $10 \times 10 = 100$ groups on this dataset. During training, the samples across input groups are imbalanced owing to the spurious correlation where every digit majorly occurs in its associated color. During test, samples across groups become balanced - thus testing a method's ability to generalize to minority groups existing in the training set.

Algorithm	**Average Accuracy**		**Worst-Group Accuracy**		**Worst-10 Group Accuracy**	
	Original	**ISR-Multiclass**	**Original**	**ISR-Multiclass**	**Original**	**ISR-Multiclass**
ERM	58.20±1.03	**78.50±0.76**	0.00±0.00	**21.93±13.40**	2.35±0.59	**39.60±7.90**
IB-ERM	70.58±1.24	**81.40±1.15**	0.94±1.07	**27.36±11.42**	10.06±2.66	**42.63±6.77**
IRM	73.85±0.79	**82.01±0.97**	8.31±2.55	**34.33±9.27**	25.66±3.14	**45.36±6.46**
IB-IRM	77.81±0.84	**82.95±2.42**	9.73±5.20	**32.29±6.66**	32.14±2.48	**49.17±5.63**

Table 1: Evaluation of ISR Multiclass on MC-CMNIST. We report the test accuracy (%) with standard deviation over 5 random trials. A value in bold indicates higher accuracy. ISR-Multiclass outperforms both average and worst group accuracies, especially for ERM. Note that the variance for worst group accuracies is high because of the less number of samples per group (≈ 300).

We report the average accuracy (across all groups), worst group accuracy, and worst-10 group accuracy (average across 10 worst groups). We compare the performance of ISR-Multiclass with ERM, IB-ERM, IRM [4] and IB-IRM which are proposed by Arjovsky et al. [2019] and Ahuja et al. [2021]. The oracle on this dataset achieves 99.03 ± 0.08 average accuracy as per Ahuja et al. [2021]. More details can be found in Appendix B.

It is evident that post-processing with ISR-Multiclass significantly improves both the average and worst-group accuracies, especially prevalent for ERM and IB-ERM. While IRM and IB-IRM perform better than their ERM counterparts, ISR-Multiclass still improves the accuracy by $\approx 5 - 10\%$. Note that ISR-Multiclass is a simple post-processing technique as compared to other methods which rely on end-to-end training. We discuss the merits of this in more detail in Appendix C.

4 Conclusion

In this work, we propose ISR-Multiclass, an extension of ISR-Mean to a more practical setting of multi-class classification. We theoretically prove that ISR-Multiclass can recover invariant features in fewer environments by using class information, specifically in $\lceil d_{spu}/k \rceil + 1$ environments. We corroborate our theory with empirical results by curating a new multi-class linear unit test and demonstrate faster recovery of invariant features by ISR-Multiclass. Further, our method outperforms current methods on the MC-CMNIST by a large margin, both for average and worst group accuracy.

[3] Note that we condition on a fixed class (0) to enable this comparison.

[4] To ensure a fair comparison, IRM was re-trained with the groups denoted by digit color.

References

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Benjamin Aubin, Agnieszka Słowik, Martin Arjovsky, Leon Bottou, and David Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Tim Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson. Finding the subspace mean or median to fit your need. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1082–1089, 2014.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2021.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. *arXiv preprint arXiv:2201.12919*, 2022.

A Appendix

A.1 Detailed Steps for ISR-Mutliclass

Step 1. Estimating sample means for every environment and every class

Construct the following matrix M_k for class k where each row contains the sample mean conditioned on a given environment e, for class k. In other words, each row is $\bar{x}_{ek} = A\mu_k + B\mu_{ke}$. Note that in the infinite sample setting considered in our work, this is exactly the mean as per Equation (4).

$$\mathcal{M}_k := \begin{bmatrix} \bar{x}_{1k}^\top \\ \vdots \\ \bar{x}_{Ek}^\top \end{bmatrix} \tag{8}$$

Step 2. Eigendecompose every M_k

For every class, eigen-decompose to obtain eigenvectors $\{P_i\}_{i\in d}^{d}$ corresponding to eigenvalues $\{\lambda_i\}_{i=1}^{d}$. By assumption 1, we obtain $E-1$ eigenvectors (-1 from the mean centering in PCA) which correspond to non-zero eigenvalues.

$$\mathcal{P}_k := \left[P_{1k}|P_{2k}|\cdots|P_{(E-1)k}\right] \in \mathbb{R}^{d\times(E-1)} \tag{9}$$

where P_{ik} is the i^{th}eigenvector corresponding to a non-zero eigenvalue in the phase transition of M_k. Thus, P_k recovers spurious dimensions corresponding to class k, as follows from ISR-Mean in Wang et al. [2022].

Note that unlike Wang et al. [2022], we *do not impose* the condition that $E > d_s$ as information from a single class may not be sufficient to recover all spurious (and thus invariant) features.

Step 3: Stack all P_k and take SVD (Singular Value Decomposition) After obtaining P_k for every class y_k, we stack all P_k to obtain a new matrix M_{total} as follows:

$$\mathcal{M}_{total} := [P_1|P_2|\cdots|P_k] \in \mathbb{R}^{d\times(E-1)k} \tag{10}$$

Next, take the SVD of P_k. Note that this step is motivated by the flag-mean Marrinan et al. [2014] to find the *common spurious subspace* for all class labels.

Step 4: Extract Spurious Feature Subspace In Theorem 1, we prove that SVD of M_{total} leads to d_s non zero singular values, where the corresponding vectors $\{P'_1, \ldots, P'_{d_s}\}$ recover the underlying spurious subspace. We stack these vectors as a matrix P'.

$$P' := [P'_1, \ldots, P'_{d_s}]^\top \in \mathbb{R}^{d_s\times d} \tag{11}$$

Step 5: Train a Classifier in the Null Space of Spurious-Feature Subspace This final step involves training a classifier in the *null space* of the extracted spurious feature subspace, which is the invariant feature subspace:

$$P'' = NullSpace(P') \in \mathbb{R}^{d\times d_c} \tag{12}$$

A.2 Proof for Theorem 1

Consider the matrix M_{total} as defined in (10):

$$\mathcal{M}_{total} := [P_1|P_2|\cdots|P_k] \in \mathbb{R}^{d\times(E-1)k} \tag{13}$$

By definition, $rank(M_{total}) \leq min(d, (E-1)\times k)$. This trivially implies that: $rank(M_{total}) \leq (E-1)\times k$.

Recall that each P_k recovers the spurious dimension specific to class k. In order to recover the underlying d_s dimensional subspace, the rank of $M_{total} = d_s$. Combining this fact with the above statement, we obtain the following inequality:

$$E - 1 \geq d_s/k \tag{14}$$

$$E \geq d_s/k + 1 \tag{15}$$

Thus, the minimum number of environments required to recover the spurious (thus invariant) feature subspace *benefits* by leveraging information from classes. The greater the number of classes, the lesser the number of training environments we require to recover the d_c dimensional invariant features. It should be noted that this decrease is observed while only leveraging the 1^{st} order moments of the class conditional data distribution.

Assuming condition (12) is satisfied, the rank of M_{total} is capped at d_s. Thus, the SVD will lead to d_s positive singular values. We can then obtain the d_s eigenvectors corresponding to these non-zero singular values, which span the spurious dimensions. The transformation matrix will be $P' \in \mathbb{R}^{d_s\times d}$. Since $z_s \perp z_c$ as per the setup 4, the null space of P' will correspond to vectors spanning the $d - d_s = d_c$ dimensions as follows:

$$NullSpace(P') = P'' \in \mathbb{R}^{d_c\times d} \tag{16}$$

Finally, training on this invariant subspace will help us obtain the optimal invariant predictor as defined in Rosenfeld et al. [2020], which completes the proof.

B Experimental Details

B.1 Multiclass Linear Unit Test

We now discuss experimental details for constructing and evaluating on Example3/3s-Multiclass, a Multi Class Linear Unit Test.

Construction First, we sample y from a multinomial distribution of uniform probability $1/k$, where k is the number of classes. Then, the first d_c invariant features are sampled from a Gaussian distribution where the mean depends on the class label. Similarly, the next d_s spurious features are sampled from a Gaussian distribution where the mean now depends on the class label *as well as* the environment label. This can be formulated as follows:

For a given environment e,

$$y \sim \text{Multinomial}\left(\frac{1}{k}\right),$$

$$z_c \sim \{\ \mathcal{N}(\mu_k, \sigma_c I_{d_c}) * \nu_{inv} \quad \text{for } y = k,$$

$$z_e \sim \{\ \mathcal{N}(\mu_{ke}, \sigma_e I_{d_s}) * \nu_{spu} \quad \text{for } y = k, env = e,$$

$$z \leftarrow \begin{bmatrix} z_c \\ z_e \end{bmatrix}, \qquad x = Rz$$

For Example3-Multiclass, $R = I_d$, and for Example3s-Multiclass (scrambled variation), $R \in \mathbb{R}^{d \times d}$ is an orthonormal matrix.

The sampling of means μ_k, μ_{ke} is done from a uniform distribution between $[0, 1)$. ν_{inv} and ν_{spu} are the scale of invariant and spurious features. We set $\mu_{inv} = 0.1$ and $\mu_{spu} = 1$ as regularization may encourage learning spurious features, making it harder to learn the invariant features. This is similar to Example2 in Aubin et al. [2021].

In our experiments, $\sigma_c = 0.1, \sigma_e = 0.1, d_s = 5, d_c = 5$. We sample 10,000 points per environment. k varies from 2 to 7.

Code and Hyperparameters For all methods, we perform a hyperparameter search over 5 data seeds and 5 model trials. In every trial, we train the algorithm on the train split and use the Adam Kingma and Ba [2014] optimizer for optimization. The model with the least mean validation error across all environments is chosen. For ISR-Mean, we use the implementation from their released code at `https://github.com/Haoxiang-Wang/ISR`.

B.2 Multiclass Coloured MNIST (MC-CMNIST)

Figure B.2 depicts a summary of the dataset - for every digit, the corresponding colour is highly correlated in the training set. This correlation breaks during testing.

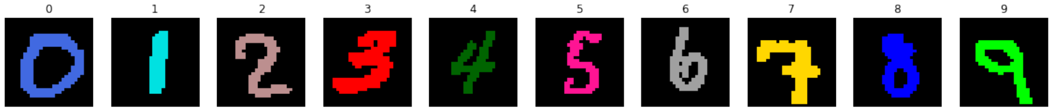


Figure 4: Multiclass Coloured MNIST dataset.

We directly employ the Multiclass coloured MNIST dataset, models and hyperparameters provided by Ahuja et al. [2021] at `https://github.com/ahujak/IB-IRM`.

For ERM, IB-ERM, IRM and IB-IRM, we run a sweep over hyperparameters using the grid as suggested above. The best model is chosen by using train domain validation (Gulrajani and Lopez-Paz [2020]). ISR-Multiclass is applied on the last-layer over the classification weights to enable the invariant feature subspace transformation. Note that ISR-Multiclass uses colour labels as the group information, and we ensure this same definition applies to IRM to ensure a fair comparison. It should also be noted that similar to Wang et al. [2022], we adopt a strategy of scaling down the spurious dimensions extracted from convolutional networks.

C Post-processing v/s end-to-end training

It should be noted that ISR-Multiclass is a *post-processing* technique and can be applied on top of *any* pretrained embeddings, while other methods rely on end-to-end training. This further demonstrates the possibility of using such a linear transformation on any embeddings obtained from large pretrained models. Fine-tuning such models maybe infeasible, and applying ISR-Multiclass (vis-a-vis linear probing) can give us the additional benefit of robustness to spurious correlations.